

DESCRIPTION LOGIC FOR DATABASES

Tesina di Intelligenza Artificiale
di Alessia Candido
Prof. Daniele Nardi

INTRODUZIONE

Logica descrittiva applicata al data management
come:

1. modellazione concettuale/semantica di sorgenti di dati
2. uso della DL nell'esprimere e valutare/processare query.
3. integrazione di diverse sorgenti in ambiente distribuito di data management

Alessia Candido

Tesina I.A.

2

CHE COS'E' UN DATABASE?

Un database è una collezione coerente di dati correlati tra loro.

Le principali nozioni relative ai databases sono:

- Schema concettuale: ER data model, o semantic modeling languages
- Schema logico: modello di dati relazionale o OO-DBMS
- Linguaggi d'interrogazione: SQL, "conjunctive" query, datalog
- Livello fisico
- Databases distribuiti

Alessia Candido

Tesina I.A.

3

DATABASE vs Knowledge Base

- I database gestiscono collezioni persistenti e di grandi dimensioni di dati relativamente semplici.
- Le basi di conoscenza interessano quantità più esigue ma più complesse di dati, inoltre permettono inferenza sui dati.
- I DBs utilizzano solitamente il data manipulation language(DML) ed il data definition language(DDL) rispettivamente per interrogare la base di dati e popolarla.
- Le DL usano le operazioni di *tell* e *ask* definite sulla KB per implementare il DDL e il DML.

Alessia Candido

Tesina I.A.

4

INTRODUZIONE

Logica descrittiva applicata al data management
come:

1. modellazione concettuale/semantica di sorgenti di dati
2. uso della DL nell'esprimere e valutare/processare query.
3. integrazione di diverse sorgenti in ambiente distribuito di data management

Alessia Candido

Tesina I.A.

5

MODELLI DI DATI E DESCRIZIONE LOGICA

Un modello di dati è essenzialmente un linguaggio o un insieme di concetti per descrivere una classe di basi di dati di un certo tipo.

Consideriamo le relazioni tra tali data models e la DL. In particolare esamineremo:

- La formalizzazione di un modello di dati: ER
- Scelta di un'appropriata DL: DLR
- La definizione della traduzione dal modello ER al DLR
- I benefici derivanti da tale traduzione

Alessia Candido

Tesina I.A.

6

MODELLO ENTITA'-RELAZIONE

Gli elementi base di tale modello a livello intenzionale (parti dello schema concettuale) sono:

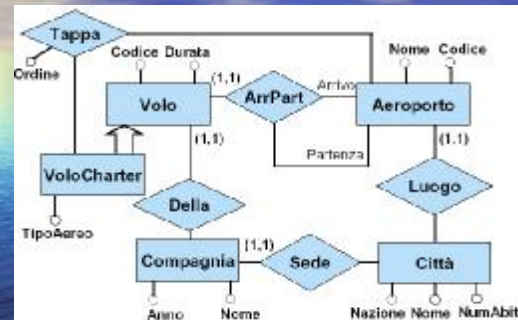
- **Entità:** classe di oggetti (istanze) con esistenza autonoma e proprietà comuni (attributi).
- **Attributo:** proprietà locale dell'entità o relazione a cui è associato univocamente un dominio D .
- **Relazione:** legame tra due o più entità; ogni entità partecipa alla relazione con un certo ruolo.
- **ISA:** relazione di sottoinsieme tra due entità.
- **Vincoli di cardinalità**

Alessia Candido

Tesina I.A.

7

ESEMPIO DI SCHEMA CONCETTUALE ER



Alessia Candido

Tesina I.A.

8

SEMANTICA DEL MODELLO ENTITA'-RELAZIONE

Uno stato di un database B , corrispondente a un ER schema, è costituito da un insieme Δ , t.c. $\Delta \cap D = \emptyset$, e una funzione $.B$ che mappa:

- **Entità:** ogni entità E in $E \subseteq \Delta$.
- **Attributo:** ogni attributo A in $A \subseteq (\Delta \times D)$, t.c. per ogni $e \in E$ esiste un solo $a = (e, d) \in A$ con $d \in D$ (vale lo stesso discorso per gli attributi di relazioni).
- **Relazione:** ogni relazione R all'insieme R di tuple etichettate di elementi di Δ , nella forma $\langle U_1: e_1, \dots, U_n: e_n \rangle$ t.c. U_i è un ruolo univoco in R , ed $e_i \in E_i$ con $i=1..n$, arità della relazione.

Alessia Candido

Tesina I.A.

9

SEMANTICA DEL MODELLO ENTITA'-RELAZIONE (2)

- **ISA:** è una relazione binaria $\leq_s$, t.c. se $E_1 \leq_s E_2$ allora $E_1 \subseteq E_2$ e se $R_1 \leq_s R_2$ allora $R_1 \subseteq R_2$.
- **Vincoli di cardinalità:** date due funzioni $cmin_s, cmax_s$ dall'insieme dei ruoli all'insieme degli interi, allora per ogni $e \in E$ si ha $cmin_s(U) \leq |\{r \in R \mid r[U_i] = e\}| \leq cmax_s(U)$.

Alessia Candido

Tesina I.A.

10

REVIEW SYNTAX OF DLR

DLR è una generalizzazione del linguaggio di DL ai ruoli n-ari:

- Concetti atomici possono essere :
 $C := T_1 \mid A \mid \neg C \mid C_1 \sqcap C_2 \mid \text{Esiste}[\$i]R \mid \leq k[\$i]R$
- Ruoli primitivi possono essere del tipo:
 $R := T_n \mid P \mid \neg R \mid R_1 \sqcap R_2 \mid (\$i/n:C)$

La forma principale di ragionamento rimane la **sussunzione** $A \sqsubseteq_s B$ cioè la verifica che B è più generale di A .

Alessia Candido

Tesina I.A.

11

REVIEW SYNTAX OF DLR(2)

In dettaglio:

- **Esiste $[\$i]R$** individua tutti gli oggetti che partecipano alla relazione R come componente i -sima (e rappresenta una proiezione nel modello relazionale);
- **$(\$i/n:C)$** individua tutte le tuple di arità n della relazione di arità n con la componente i -esima oggetto del concetto C (e rappresenta una selezione nel modello relazionale);
- **$\leq k[\$i]R$** è una restrizione numerica sulla relazione (e rappresenta i vincoli di cardinalità sui ruoli nel modello relazionale);

Alessia Candido

Tesina I.A.

12

TRADUZIONE DAL ER-SCHEMA AL DLR-KB

- L'insieme delle entità corrispondono all'insieme dei concetti atomici, e ogni istanza di entità corrisponde ad un oggetto del concetto in modo univoco.
- Ogni relazione di arità n corrisponde ad un ruolo primitivo di arità n rappresentato da P_R .
- Ogni attributo corrisponde ad un ruolo di arità 2 rappresentato da P_A , cioè ogni oggetto di P_A è una coppia dove il primo componente è l'istanza dell'entità ed il secondo è un elemento del concetto corrispondente al dominio dell'attributo.

Alessia Candido

Tesina I.A.

13

TRADUZ. DAL ER-SCHEMA AL DLR-KB(2)

- Per ogni E_1, E_2 t.c. $E_1 \leq_s E_2$ vale $E_1 C_s E_2$
- Per ogni R_1, R_2 t.c. $R_1 \leq_s R_2$ vale $P_{R_1} C_s P_{R_2}$
- Per ogni $A \subseteq (ExD)$, vale $E C_s (\text{Per ogni}[\$1](P_A \Pi (\$2:C))) \Pi = 1[\$1]P_A$

Essendo μ_R la funzione che dato un ruolo U della relazione ER restituisce l'intero corrispondente alla posizione i-esima che il ruolo occupa nella tupla, e presa R relazione di arità n con ruoli U_i

- $P_R C_s (\$ \mu_R(U_1): E_1) \Pi \dots \Pi (\$ \mu_R(U_n): E_n)$
- Se $m = cmin_s(U_i)$ e $m \neq 0$ vale $E_i C_s \geq m[\$ \mu_R(U_i)]P_R$
- Se $n = cmax_s(U_i)$ e $n \neq \infty$ vale $E_i C_s \leq n[\$ \mu_R(U_i)]P_R$

Alessia Candido

Tesina I.A.

14

TRADUZIONE DELL'ESEMPIO ER

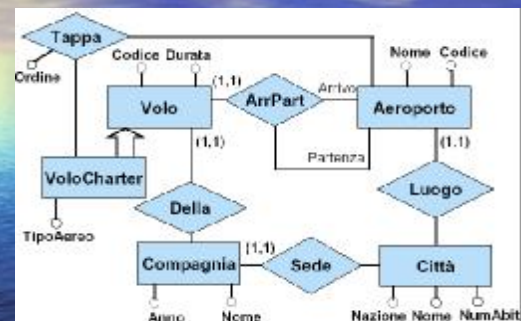
- **Tappa** $C_s(\$1:\text{VoloCharter}) \Pi (\$2:\text{Aeroporto}) \Pi \text{Per ogni}[\$1](\text{Ordine} \Pi (\$2:\text{int})) \Pi = 1[\$1]\text{Ordine}$
- **Arr-part** $C_s(\$1:\text{Volo}) \Pi (\$ \mu_R(\text{arrivo}): \text{Aeroporto}) \Pi (\$ \mu_R(\text{partenza}): \text{Aeroporto}) \Pi = 1[\$1]\text{Arr-part}$
- **Della** $C_s(\$1:\text{Volo}) \Pi (\$2:\text{Compagnia}) \Pi = 1[\$1]\text{Della}$
- **Sede** $C_s(\$1:\text{Compagnia}) \Pi (\$2:\text{Città}) \Pi = 1[\$1]\text{Sede}$
- **Luogo** $C_s(\$1:\text{Aeroporto}) \Pi (\$2:\text{Città}) \Pi = 1[\$1]\text{Luogo}$
- Compagnia C_s Per ogni $[\$1](\text{Anno} \Pi (\$2:\text{int})) \Pi = 1[\$1]\text{Anno}$ Per ogni $[\$1](\text{Nome} \Pi (\$2:\text{string})) \Pi = 1[\$1]\text{Nome}$

Alessia Candido

Tesina I.A.

15

ESEMPIO DI SCHEMA CONCETTUALE ER



Alessia Candido

Tesina I.A.

16

TRADUZIONE DELL'ESEMPIO ER (2)

- Volo C_s Per ogni $[\$1](\text{Codice} \Pi (\$2:\text{int})) \Pi = 1[\$1]\text{Codice}$ Per ogni $[\$1](\text{Durata} \Pi (\$2:\text{time})) \Pi = 1[\$1]\text{Durata}$
- Aeroporto C_s Per ogni $[\$1](\text{Codice} \Pi (\$2:\text{int})) \Pi = 1[\$1]\text{Codice}$ Per ogni $[\$1](\text{Nome} \Pi (\$2:\text{string})) \Pi = 1[\$1]\text{Nome}$
- Città C_s Per ogni $[\$1](\text{Nazione} \Pi (\$2:\text{string})) \Pi = 1[\$1]\text{Nazione}$ Per ogni $[\$1](\text{NumAbit} \Pi (\$2:\text{int})) \Pi = 1[\$1]\text{NumAbit}$ Per ogni $[\$1](\text{Nome} \Pi (\$2:\text{string})) \Pi = 1[\$1]\text{Nome}$
- VoloCharter C_s Volo Π Per ogni $[\$1](\text{TipoAereo} \Pi (\$2:\text{int})) \Pi = 1[\$1]\text{TipoAereo}$

Alessia Candido

Tesina I.A.

17

VANTAGGI DELLA TRADUZIONE

Alcuni esempi di funzionalità aggiunte al modello ER da questa traduzione sono:

- **Generalizzazione delle gerarchie:**
la possibilità di esprimere un'entità come unione disgiunta di altre entità, usando la negazione e l'unione del DLR.
- **Raffinamenti delle proprietà:**
la possibilità di specificare le proprietà di entità lungo una gerarchia di ISA, piuttosto che una semplice aggiunta di attributi, e di imporre vincoli di cardinalità per restringere anche la gamma di valori di una certa proprietà con l'uso dei quantificatori.
- **Definizione di classi di oggetti:**
attraverso le condizioni necessarie e sufficienti da soddisfare, senza esplicitamente istanziare gli oggetti.

Alessia Candido

Tesina I.A.

18

VANTAGGI DELLA TRADUZIONE(2)

- **Vincoli di chiave:**

Calvanese ha mostrato come il ragionamento sulla DLR con tali vincoli non peggiora la complessità computazionale nel caso peggiore.

- **Vincoli temporali:**

la possibilità di catturare le informazioni variate nel tempo attraverso un'estensione del DLR. Artale ha studiato la complessità e decidibilità di tale estensione chiamata DLRus.

Alessia Candido

Tesina I.A.

19

VANTAGGI DELLA TRADUZIONE(3)

Inoltre la formalizzazione dello schema ER in termini della DLR permette di applicare le seguenti forme di ragionamento alla base di dati:

- **Soddisfacibilità di entità, relazioni e consistenza dello schema ER:**

verifica se rispettivamente i concetti, i ruoli e l'intera base di conoscenza dallo schema dedotta ammette un modello. In particolare lo schema ER si dice inconsistente se la totalità delle definizioni nella KB sono inconsistenti o ammettono un modello infinito.

- **Verifica di non ridondanza nello schema.**
- **Sussunzione tra entità e relazioni.**

Alessia Candido

Tesina I.A.

20

INTRODUZIONE

Logica descrittiva applicata al data management come:

1. modellazione concettuale/semantica di sorgenti di dati
2. **uso della DL nell'esprimere e valutare/processare query.**
3. integrazione di diverse sorgenti in ambiente distribuito di data management

Alessia Candido

Tesina I.A.

21

INTERROGARE BASE DI DATI IN DL

Poiché abbiamo visto che le definizioni dei concetti possono essere considerate come un insieme di condizioni necessarie e sufficienti che gli oggetti del concetto devono soddisfare, una query può essere trattata come una definizione di concetto. Tale, ad esempio, è la scelta adottata dal sistema **Classic**: unificare in un unico linguaggio il data definition e data manipulation language.

Alessia Candido

Tesina I.A.

22

INTERROGARE BASE DI DATI IN DL(2)

Partendo da questa osservazione

- **query=concept description**

Osserviamo che a favore di una maggiore efficienza ed espressività, c'è anche una maggiore complessità del DL e quindi una effettiva difficoltà da parte dell'utente a porre interrogazioni alla KB.

Per risolvere il problema di query inconsistenti sono state proposte diverse soluzioni, basate più o meno tutte sui seguenti principi:

- **Interrogazioni iterative**, a passi successivi di raffinamento
- **Caching** di query più frequenti e **classificazione** all'interno della gerarchia di sussunzione

Alessia Candido

Tesina I.A.

23

QUERY ESTESE

Uno dei limiti della DL risiede nel fatto che le query esprimibili possono ritornare solo sottoinsiemi di oggetti preesistenti e non crearne di nuovi.

Due sono gli approcci a questo problema perseguiti, in entrambi i casi estendendo le query standard:

1. Le query sono trattate come vincoli di tipo sulle variabili presenti in formule di Horn. Ciò comporta una netta distinzione tra le definizioni dei concetti e ruoli e le relazioni usate per esprimere regole sulla KB.
2. I concetti e i ruoli sono definiti come predicati unari e binari che possono apparire come atomi di una query. Ciò permette di creare query congiunzione di altre query già definite sulla KB.

Alessia Candido

Tesina I.A.

24

QUERY ESTESE(2)

Un'altra distinzione importante sui tipi di query che una base di dati in DLR può supportare è:

- **Query non ricorsive** (QNR)
- **Query ricorsive** (QR)

Nel caso di QNR il sistema d'interrogazione rimane **decidibile**.

La scelta di rendere le query ricorsive, o di aggiungere espressività al DL (con l'uso della negazione, dei restrittori numerici, della disgiunzione..) comporta che il problema di rispondere alle interrogazioni della KB diventi velocemente **indecidibile**.

Alessia Candido

Tesina I.A.

25

OTTIMIZZAZIONE SEMANTICA DI QUERY

- [Beck] **Classifica** la query nello schema dei concetti, e applica la valutazione della query ai soli oggetti del concetto da cui è sussunto.
- **Memorizza** le viste risultato delle query precedenti in ordine di sussunzione e per ogni nuova query procedi come al p.to precedente.
- [Buchheit] **Query=structural part+dirty part**

La parte strutturale è quella espressa in DL e la vista ottenuta applicando solo questa parte alla KB è cached.

La parte sporca è in logica del primo ordine ed è applicata alla vista cached per restituire la risposta alla query.

- [Bergamaschi] **Eliminazione delle ridondanze**

La query viene espansa, cioè diventa la congiunzione di tutte le regole da cui è sussunta.

Dalla query espansa si eliminano i termini ridondanti e si esegue.

La vista ottenuta si classifica e si memorizza nella gerarchia preesistente.

Alessia Candido

Tesina I.A.

26

INTRODUZIONE

Logica descrittiva applicata al data management come:

1. modellazione concettuale/semantica di sorgenti di dati
2. uso della DL nell'esprimere e valutare/processare query.
3. **Integrazione di diverse sorgenti in ambiente distribuito di data management**

Alessia Candido

Tesina I.A.

27

DATA INTEGRATION

L'obiettivo del data integration system (DIS) è di fornire un'interfaccia uniforme per multiple sorgenti di dati.

La DL in quest'ambito ha due ruoli:

- Specificare il contenuto delle varie sorgenti:
 - a livello concettuale
 - a livello logico
- Processare le interrogazioni poste al sistema integrato sulla base delle specifiche delle sorgenti.

Alessia Candido

Tesina I.A.

28

SPECIFICA A LIVELLO CONCETTUALE

Il livello concettuale esprime la semantica dell'applicazione. Distinguiamo i seguenti elementi di un DIS:

- Enterprise Conceptual Schema (ECS)
- Source Conceptual Schema (SCS)
- Domain Conceptual Schema (DCS)
dove il DCS=ECS+SCS+ISR(inter-schema-relationship)

Alessia Candido

Tesina I.A.

29

SPECIFICA A LIVELLO CONCETTUALE (2)

Per rappresentare gli schemi concettuali del DIS adotteremo nuovamente il DLR, mentre per definire le ISR faremo uso delle seguenti asserzioni in DL:

- $L_i \text{Cext } L_j$:
ogni oggetto che soddisfa l'espressione L_i nella sorgente i soddisfa anche l'espressione L_j nella sorgente j , cioè L_i è estenzionalmente incluso in L_j .
- $L_i \text{Cint } L_j$:
il concetto denotato dall'espressione L_i nella sorgente i è un sottoconcetto di uno denotato dall'espressione L_j nella sorgente j .

Alessia Candido

Tesina I.A.

30

SPECIFICA A LIVELLO LOGICO

Il livello logico fornisce una descrizione di quale modello di DBMS è adottato da ogni sorgente. Il link tra questo livello e il DCS è implementato in due modi:

- **Approccio a visione globale:** ogni query è associata a un concetto nel DCS, cioè il concetto è trattato come una vista sopra le sorgenti.
- **Approccio a visione locale:** ad ogni relazione di ogni sorgente è associata una query nei termini del DCS.

Alessia Candido

Tesina I.A.

31

APPROCCIO A VISIONE LOCALE

Per descrivere il contenuto delle sorgenti attraverso viste definiamo una relazione di una sorgente nei termini di una query, nel seguente modo:

$$q(\underline{x}) \leftarrow conj_1(\underline{x}, \underline{y}_1) \vee \dots \vee conj_m(\underline{x}, \underline{y}_m)$$

- La testa $q(\underline{x})$ è la definizione della relazione con nome e numero di colonne.
- Il corpo descrive il contenuto della relazione come unione di query:
 $conj_m(\underline{x}, \underline{y}_m)$ è la congiunzione di atomi (entità, relazioni e attributi del DCS) con $\underline{x}$ e $\underline{y}_m$ le variabili presenti nella congiunzione.

Alessia Candido

Tesina I.A.

32

INFERENZA SUL DIS

Questa rappresentazione del DIS attraverso DL richiede nuovi algoritmi per risolvere i seguenti quesiti:

- **Query containment:** date due query q_1 e q_2 , verifica se q_1 è contenuta in q_2 per ogni modello che soddisfa il DCS.
- **Query consistency:** verifica se una query è consistente per ogni modello che soddisfa il DCS.
- **Query disjointness:** verifica se due query sono disgiunte per ogni modello che soddisfa il DCS.

Alessia Candido

Tesina I.A.

33

INTERROGARE IL DIS

L'obiettivo è quello di poter interrogare il DCS e quindi in maniera trasparente ogni sorgente del sistema. I modi per processare le query basate sulle viste, secondo un approccio a visione locale, sono:

- **Query rewriting:** le query vengono riformulate in accordo con le viste di ogni sorgente coinvolta e poi sono valutate singolarmente.
- **Query answering:** si estendono le viste e si valutano tutte le tuple che soddisfano la query in tutte le basi di dati consistenti con le viste.

Ogni soluzione al problema del processamento di query basate sulle viste è un'elaborazione accurata di tali approcci.

Alessia Candido

Tesina I.A.

34

CONCLUSIONI

Vantaggi:

- La DL ha la capacità di rappresentare la semantica di qualsiasi sorgente di dati
- Ha una semantica chiara soprattutto non ambigua e precisa (in particolare rispetto a modelli di DB semistrutturati)
- Può rappresentare conoscenza ma anche fare inferenza su di essa, attraverso la sussunzione
- Permette di verificare la consistenza delle definizioni sulla base di dati e dell'intero modello
- Ottimizza il processamento delle interrogazioni e la valutazione del risultato è decidibile con poche restrizioni.

Alessia Candido

Tesina I.A.

35

CONCLUSIONI(2)

Limiti:

- La DL, a causa della sua complessità espressiva, non ha raggiunto il successo e la diffusione dei database relazionali.
- La DL standard non permette di generare nuovi oggetti nella base di dati ma solo selezionare sottoinsiemi di oggetti preesistenti.

Alessia Candido

Tesina I.A.

36